Mapping-based genome size estimation

Boas Pucker¹,²*

1 Genetics and Genomics of Plants, Bielefeld University, Bielefeld, Germany
2 Center for Biotechnology (CeBiTec); Bielefeld University, Bielefeld, Germany

Email: bpucker@cebitec.uni-bielefeld.de

ORCID: 0000-0002-3321-7471

* corresponding author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

Key words: NGS, genome sequencing, k-mer profile, comparative genomics, Arabidopsis thaliana, Beta vulgaris, Solanum lycopersicum, Brachypodium distachyon, Vitis vinifera, Zea mays

Abstract

While the size of chromosomes can be measured under a microscope, the size of genomes cannot be measured precisely. Biochemical methods and k-mer distribution-based approaches allow only estimations. An alternative approach to predict the genome size based on high contiguity assemblies and short read mappings is presented here and optimized on Arabidopsis thaliana and Beta vulgaris. Brachypodium distachyon, Solanum lycopersicum, Vitis vinifera, and Zea mays were also analyzed to demonstrate the broad applicability of this approach. Mapping-based Genome Size Estimation (MGSE) and additional scripts are available on github: https://github.com/bpucker/MGSE.
Introduction

Nearly all parts of the plant are now tractable to measure, but assessing the size of a plant genome is still challenging. Although chromosome sizes can be measured under a microscope [1], the combined length of all DNA molecules in a single cell is still unknown. Almost 20 years after the release of the first *Arabidopsis thaliana* genome sequence, this holds even true for one of the most important model species. Initially, biochemical methods like reassociation kinetics [2], Feulgen photometry [3], quantitative gel blot hybridization [4], southern blotting [5], and flow cytometry [6, 7] were applied. Unfortunately, these experimental methods rely on a reference genome [8]. The rise of next generation sequencing technologies [9] enabled new approaches based on k-mer profiles or the counting of unique k-mers [10, 11]. JellyFish [11], Kmergenie [12], Tallymer [13], Kmerlight [14], and genomic character estimator (gce) [15] are dedicated tools to analyze k-mers in reads. Next, genome sizes can be estimated based on unique k-mers or a complete k-mer profile. Many assemblers like SOAPdenovo [16] and ALLPATHS-LG [17] perform an internal estimation of the genome size to infer an expected assembly size. Recently, dedicated tools for the genome size estimation like GenomeScope [18] and findGSE [19] were developed. Although the authors considered and addressed a plethora of issues with real data [18], results from different sequencing data sets for the same species can vary. While some proportion of this variation can be attributed to accession-specific differences as described e.g. for *A. thaliana* [19, 20], specific properties of a sequencing library might have an impact on the estimated genome size. For example, high levels of bacterial or fungal contamination could bias the result if not removed prior to the estimation process. Due to high accuracy requirements, k-mer-based approaches are usually restricted to high quality short reads and cannot be applied to long reads of third generation sequencing technologies. The rapid development of long read sequencing technologies enables high contiguity assemblies for almost any species and is therefore becoming the standard for genome sequencing projects [21, 22]. Nevertheless, some highly repetitive regions of plant genomes like nucleolus organizing region (NOR) and centromeres remain usually unassembled [20, 23, 24]. Therefore, the genome size cannot be inferred directly from the assembly size, but the assembly size can be considered a lower boundary when estimating genome sizes.

Extreme genome size estimates of *A. thaliana* for example 70 Mbp [2] or 211 Mbp [25] have been proven to be inaccurate based on insights from recent assemblies [20, 24, 26–28]. However, various methods still predict genome sizes between 125 Mbp and 165 Mbp for diploid *A. thaliana* accessions [26, 29–31]. Substantial technical variation is observed not only between
methods, but also between different labs or instruments [32]. As described above, extreme
examples for A. thaliana display 3 fold differences with respect to the estimated genome size.
Since no assembly is representing the complete genome, the true genome size remains
unknown. An empirical approach, i.e. running different tools and comparing the results, might be
a suitable strategy.

This work presents a method for the estimation of genome sizes based on the mapping of reads
to a high contiguity assembly. Mapping-based Genome Size Estimation (MGSE) is a Python
script which processes the coverage information of a read mapping and predicts the size of the
underlying genome. MGSE is an orthogonal approach to the existing tools for genome size
estimation with different challenges and advantages.

Methods

Data sets

Sequencing data sets of the A. thaliana accessions Columbia-0 (Col-0) [33–38] and
Niederzenz-1 (Nd-1) [31] as well as several Beta vulgaris accessions [39–41] were retrieved
from the Sequence Read Archive (AdditionalFile 1). Only the paired-end fraction of the two
included Nd-1 mate pair libraries was included in this analysis. Genome assembly versions
TAIR9 [42], AthNd-1_v1 [31], AthNd-1_v2 [24], and RefBeet v1.5 [39, 43] served as references
in the read mapping process. The A. thaliana assemblies, TAIR9 and Ath-Nd-1_v2, already
included plastome and chondrome sequences. These subgenome sequences of Ath-Nd-1_v2
were added to Ath-Nd-1_v1 as this assembly was previously cleaned of such sequences.
Plastome (KR230391.1, [44]) and chondrome (BA000009.3, [45]) sequences were added to
RefBeet v1.5 to allow proper placement of respective reads.

Genome sequences of Brachypodium distachyon strain Bd21 (GCF_000005505.3 [46]),
Solanum lycopersicum (GCA_002954035.1 [47]), Vitis vinifera cultivar Chardonnay
(QGNW01000001.1 [48]), and Zea mays cultivar DK105 (GCA_003709335.1 [49]) were
retrieved from the NCBI. Corresponding read data sets were retrieved from the Sequence Read
Archive (AdditionalFile1).
Genome size estimation

JellyFish2 v2.2.4 [11] was applied for the generation of k-mer profiles which were subjected to GenomeScope [18]. Selected k-mer sizes ranged from 19 to 25. Results of different sequencing data sets and different k-mer sizes per accession were compared. Genomic character estimator (gce) [15] and findGSE [19] were applied to infer genome sizes from the k-mer histograms. If tools failed to predict a value or if the prediction was extremely unlikely, values were masked to allow meaningful comparison and accommodation in one figure. The number of displayed data points is consequently a quality indicator.

Mapping-based genome size estimation

Despite some known biases [50–52], the underlying assumption of MGSE is a nearly random fragmentation of the DNA and thus an equal distribution of sequencing reads over the complete sequence. If the sequencing coverage per position (C) is known, the genome size (N) can be calculated by dividing the total amount of sequenced bases (L) by the average coverage value: \( N = \frac{L}{C} \). Underrepresented repeats and other regions display a higher coverage, because reads originating from different genomic positions are mapped to the same sequence. The accurate identification of the average coverage is crucial for a precise genome size calculation. Chloroplastic and mitochondrial sequences account for a substantial proportion of reads in sequencing data sets, while contributing very little size compared to the nucleome. Therefore, sequences with very high coverage values i.e. plastome and chondrome sequences are included during the mapping phase to allow correct placement of reads, but are excluded from MGSE. A user provided list of reference regions is used to calculate the median or mean coverage based on all positions in these specified regions. Benchmarking Universal Single Copy Orthologs (BUSCO) [53] can be deployed to identify such a set of \textit{bona fide} single copy genes which should serve as suitable regions for the average coverage calculation. Since BUSCO is frequently applied to assess the completeness of a genome assembly, these files might be already available to users. GFF files generated by BUSCO can be concatenated and subjected to MGSE. As some BUSCOs might occur with more than one copy, MGSE provides an option to reduce the predicted gene set to the actual single copy genes among all identified BUSCOs.
BWA MEM v0.7 [54] was applied for the read mapping and MarkDuplicates (Picard tools v2.14) [55] was used to filter out reads originating from PCR duplicates. Next, a previously described Python script [56] was deployed to generate coverage files, which provide information about the number of aligned sequencing reads covering each position of the reference sequence. Finally, MGSE (https://github.com/bpucker/MGSE) was run on these coverage files to predict genome sizes independently for each data set.

**Results & Discussion**

*Arabidopsis thaliana* genome size

MGSE was deployed to calculate the genome size of the two *A. thaliana* accessions Col-0 and Nd-1 (Fig. 1). In order to identify the best reference region set for the average coverage calculation, different reference region sets were tested. Manually selected single copy genes, all protein encoding genes, all protein encoding genes without transposable element related genes, only exons of these gene groups, and BUSCOs were evaluated (AdditionalFile2). The results were compared against predictions from GenomeScope, gce, and findGSE for k-mer sizes 19, 21, 23, and 25.

Many estimations of the Col-0 genome size are below the assembly size of 120 Mbp [26] and display substantial variation between samples (Fig. 1a). Due to low variation between different samples and a likely average genome size the BUSCO-based approaches appeared promising. GenomeScope predicted a similar genome size, while gce reported consistently much smaller values. findGSE predicted on average a substantially larger genome size. Final sample sizes below six indicated that prediction processes failed e.g. due to insufficient read numbers.

The variation among the estimated genome sizes of Nd-1 was smaller than the variation between the Col-0 samples (Fig. 1). BUSCO-based estimations differed substantially between mean and median with respect to the variation between samples (Fig. 1b). Therefore, the average coverage is probably more reliably calculated via mean than via median. While gce predicted as reasonable genome size for Nd-1, the average predictions by GenomeScope and findGSE are very unlikely, as they contradict most estimations of *A. thaliana* genome sizes [6, 19, 24, 31].
Fig. 1: Comparison of *Arabidopsis thaliana* genome size estimations.

Genome sizes of the *A. thaliana* accessions Col-0 (a) and Nd-1 (b) were predicted by MGSE, GenomeScope, gce, and findGSE. Different MGSE approaches were evaluated differing by the set of
regions for the average coverage calculation (e.g. all genes) and the methods for the calculation of this value (mean/median). Multiple read data sets (n) were analyzed by each tool/approach to infer an average genome size given as median (m, yellow line) and mean (green triangles). Transposable elements = TE, without = wo.

The genome size estimation of about 139 Mbp inferred for Nd-1 through integration of all analyses is slightly below previous estimations of about 146 Mbp [31]. Approximately 123.5 Mbp are assembled into pseudochromosomes which do not contain complete NORs or centromeric regions [24]. Based on the read coverage of the assembled 45S rDNA units, the NORs of Nd-1 are expected to account for approximately 2-4 Mbp [31]. Centrometric repeats which are only partially represented in the genome assembly [24] account for up to 11 Mbp [31]. In summary, the Nd-1 genome size is expected to be around 138-140 Mbp. The BUSCOs which occur actually with a single copy in Ath-Nd1_v2 emerged as the best set of reference regions for MGSE.

The relevance of very high assembly contiguity was assessed by comparing results of AthNd-1_v1 (AdditionalFile3), which is based on short Illumina reads, to results of AthNd-1_v2 (AdditionalFile2), which is based on long Single Molecule Real Time sequencing (PacBio) reads. The genome size predictions based on AthNd-1_v2 were substantially more accurate. Reads are not mapped to the ends of contigs or scaffolds. This has only a minor influence on large contigs, because a few small regions at the ends with lower coverage can be neglected. However, the average coverage of smaller contigs might be biased as the relative contribution of contig ends weights stronger. In addition, the representation of centrometric repeats and transposable elements increases with higher assembly size and contiguity [24].

The feasibility of MGSE was further demonstrated by estimating the genome sizes of 1,028 *A. thaliana* accessions (Fig. 2, AdditionalFile4) which were analyzed by re-sequencing as part of the 1001 genome project [57]. Most predictions by MGSE are between 120 Mbp and 160 Mbp, while all other tools predict most genome sizes between 120 Mbp and 200 Mbp with some outliers showing very small or very large genome sizes. MGSE differs from all three tools when it comes to the number of failed or extremely low genome size predictions. All k-mer-based approaches predicted genome sizes below 50 Mbp, which are most likely artifacts. This comparison revealed systematic differences between findGSE, gce, and GenomeScope with respect to the average predicted genome size. findGSE tends to predict larger genome sizes.
than gce and GenomeScope. Very large genome sizes could have biological explanations like polyploidization events.

**Fig. 2: Genome size estimations of Arabidopsis thaliana accessions.**

MGSE, findGSE, gce, and GenomeScope were deployed to predict the genome sizes of 1,028 A. thaliana accessions based on sequence read data sets (AdditionalFile4). Extreme outliers above 200 Mbp (MGSE) or 300 Mbp (other tools) are displayed at the plot edge to allow accommodation of all data points with sufficient resolution in the center.
Different sequencing data sets of *Beta vulgaris* were analyzed via MGSE, GenomeScope, gce, and findGSE to assess the applicability to larger and more complex genomes (Fig. 3, AdditionalFile5). Different cultivars served as material source for the generation of the analyzed read data sets. Therefore, minor differences in the true genome size are expected. Moreover, sequence differences like single nucleotide variants, small insertions and deletions, as well as larger rearrangements could influence the outcome of this analysis. Since the current RefBeet v1.5 assembly represents 567 Mbp [39, 43] of the genome, all estimations below this value can be discarded as erroneous. Therefore, the mean-based approaches relying on all genes or just the BUSCOs as reference region for the sequencing coverage estimation outperformed all other approaches (Fig. 3). When comparing the *A. thaliana* and *B. vulgaris* analyses, the calculation of an average coverage in all BUSCOs, which are actually present as a single copy in the investigated genome, appears to be the most promising approach. While GenomeScope and gce underestimate the genome size, the predictions by findGSE are extremely variable but mostly around the previously estimated genome sizes [39, 43]. Based on results from the *A. thaliana* investigation, the mean calculation among all single copy BUSCOs should be the best approach. The prediction of slightly less than 600 Mbp is probably an underestimation, but still the highest reliable estimate. When assuming centromere sizes of only 2-3 Mbp per chromosome, this number could be in a plausible range. However, a previous investigation of the repeat content indicates a larger genome size due to a high number of repeats which are not represented in the assembly [58].
Fig. 3: Comparison of Beta vulgaris genome size estimations.

The genome size of B. vulgaris was predicted by MGSE, GenomeScope, gce, and findGSE. Different MGSE approaches were evaluated differing by the set of regions for the average coverage calculation (e.g. all genes) and the methods for the calculation of this value (mean/median). Multiple read data sets (n) were analyzed by each tool and approach to infer an average genome size given as median (m, yellow line) and mean (green triangles).

Application to broad taxonomic range of species

After optimization of MGSE on A. thaliana (Rosids) and B. vulgaris (Caryophyllales), the tool was deployed to analyze data sets of different taxonomic groups thus demonstrating broad applicability. Brachypodium distachyon was selected as representative of grasses. Solanum lycopersicum represents the Asterids, Zea mays was included as monocot species with high transposable element content in the genome, and Vitis vinifera was selected due to a very high heterozigosity. The predictions of MGSE are generally in the same range as the predictions generated by GenomeScope, gce, and findGSE (AdditionalFile5, AdditionalFile6, AdditionalFile7, AdditionalFile8, and AdditionalFile9). With an average prediction of 290 Mbp as
The genome size of *B. distachyon*, the MGSE prediction is slightly exceeding the assembly size. GenomeScope and gce predict genome sizes below the assembly size, while the prediction of 303 Mbp by findGSE is more reasonable. The *Z. mays* genome size is underestimated by all four tools. However, MGSE outperforms GenomeScope and gce on the analyzed data set. The *S. lycopersicum* genome size is underestimated by MGSE on most data sets. However, the compared tools failed to predict a genome size for multiple read data sets. The highest MGSE predictions are in the range of the expected genome size. MGSE failed for *V. vinifera* by predicting only 50 Mbp. The high heterozigosity of this species could contribute to this by causing lower mapping rates outside of important protein encoding genes i.e. BUSCO genes.

**Considerations about performance and outlook**

MGSE performs best on a high contiguity assembly and requires a (short) read mapping to this assembly. Accurate coverage calculation for each position in the assembly is important and contigs display artificially low coverage values towards the ends. This is caused by a reduction in the number of possible ways reads can cover contig ends. The shorter a contig, the more is the apparent coverage of this contig reduced. Since a read mapping is required as input, MGSE might appear less convenient than classical k-mer-based approaches at first look. However, these input files are already available for many plant species, because such mappings are part of the assembly process [23, 24, 59, 60]. Future genome projects are likely to generate high continuity assemblies and short read mappings in the polishing process.

One advantage of MGSE is the possibility to exclude reads originating from contaminating DNA even if the proportion of such DNA is high. Unless reads from bacterial or fungal contaminations were assembled and included in the reference sequence, the approach can handle such reads without identifying them explicitly. This is achieved by discarding unmapped reads from the genome size estimation. MGSE expects a high contiguity assembly and assumes all single copy regions of the genome are resolved and all repeats are represented by at least one copy. Although the amount of contamination reads is usually small, such reads are frequently observed due to the high sensitivity of next generation sequencing [31, 61–64].

Reads originating from PCR duplicates could impact k-mer profiles and also predictions based on these profiles if not filtered out. After reads are mapped to a reference sequence, read pairs originating from PCR duplicates can be identified and removed based on identical start and end
positions as well as identical sequences. This results in the genome size prediction by GMSE being independent of the library diversity. If the coverage is close to the read length or the length of sequenced fragments, reads originating from PCR duplicates cannot be distinguished from \textit{bona fide} identical DNA fragments. Although MGSE results get more accurate with higher coverage, after exceeding an optimal coverage the removal of apparent PCR duplicates could become an issue. Thus, a substantially higher number of reads originating from PCR-free libraries could be used if duplicate removal is omitted. Depending on the sequencing library diversity completely skipping the PCR duplicate removal step might be an option for further improvement. As long as these PCR duplicates are mapped equally across the genome, MGSE can tolerate these artifacts.

All methods are affected by DNA of the plastome and chondrome integrated into the nuclear chromosomes \cite{65, 66}. K-mers originating from these sequences are probably ignored in many k-mer-based approaches, because they appear to originate from the chondrome or plastome i.e. k-mers occur with very high frequencies. The apparent coverage in the mapping-based calculation is biased due to high numbers of reads which are erroneously mapped to these sequences instead of the plastome or chondrome sequence.

Differences in the GC content of genomic regions were previously reported to have an impact on the sequencing coverage \cite{67, 68}. Both, extremely GC-rich and AT-rich fragments, respectively, are underrepresented in the sequencing output mainly due to biases introduced by PCR \cite{69, 70}. Sophisticated methods were developed to correct coverage values based on the GC content of the underlying sequence \cite{70–72}. The GC content of genes selected as reference regions for the coverage estimation is likely to be above the 36.3% average GC content of plants \cite{56}. This becomes worse when only exons are selected due to the even higher proportion of coding sequence. Although a species specific codon usage can lead to some variation, constraints of the genetic code determine a GC content of approximately 50% in coding regions. The selection of a large set of reference regions with a GC content close to the expected overall GC content of a genome would be ideal. However, the overall GC content is unknown and cannot be inferred from the reads due to the above mentioned sequencing bias. As a result, the average sequencing coverage could be overestimated leading to an underestimation of the genome size. Future investigations are necessary to develop a correction factor for this GC bias of reads.
Many plant genomes pose an additional challenge due to recent polyploidy or high heterozygosity. Once high contiguity long read assemblies become available for these complex genomes, a mapping based approach is feasible. As long as the different haplophases are properly resolved, the assessment of coverage values should reveal a good estimation of the genome size. Even the genomes of species which have recently undergone polyploidization could be investigated with moderate adjustments to the workflow. Reference regions need to be selected to reflect the degree of ploidy in their copy number.

The major issue when developing tools for the genome size prediction is the absence of a gold standard. Since as of yet there is no completely sequenced plant genome, benchmarking with real data cannot be perfect. As a result, how various estimation approaches will compare to the first completely sequenced and assembled genome remains speculative. Although not evaluated in this study, we envision that MGSE could be generally applied to all species and is not restricted to plants.

**Data availability**

Scripts developed as part of this work are freely available on github: https://github.com/bpucker/MGSE (https://doi.org/10.5281/zenodo.2636733). Underlying data sets are publicly available at the NCBI and SRA, respectively.

**Acknowledgements**

Members of Genetics and Genomics of Plants contributed to this work by discussion of preliminary results. Many thanks go to Hanna Schilbert, Nathanael Walker-Hale, and Iain Place for helpful comments on the manuscript.

**References**


8. Bennett MD, Leitch IJ, Price HJ, Johnston JS. Comparisons with Caenorhabditis (∼100 Mb) and Drosophila (∼175 Mb) Using Flow Cytometry Show Genome Size in Arabidopsis to be ∼157 Mb and thus ∼25 % Larger than the Arabidopsis Genome Initiative Estimate of ∼125 Mb. Ann Bot. 2003;91:547–57.


Supplements:

AdditionalFile1: Sequencing data set overview.

AdditionalFile2: A. thaliana genome size prediction values for all different approaches.

AdditionalFile3: A. thaliana genome size prediction based on Ath-Nd1_v1.

AdditionalFile4: A. thaliana genome size predictions by MGSE, findGSE, gce, and GenomeScope.
AdditionalFile5: *B. vulgaris*, *Zea mays*, *Brachypodium distachyon*, *Solanum lycopersicum*, and *Vitis vinifera* genome size prediction values for all different approaches.

AdditionalFile6: Genome size estimation of *Brachypodium distachyon*.

AdditionalFile7: Genome size estimation of *Zea mays*.

AdditionalFile8: Genome size estimation of *Solanum lycopersicum*.

AdditionalFile9: Genome size estimation of *Vitis vinifera*.